DOCUMENT RESUME

ED 292 877 TM 011 328

AUTHOR Woodruff, David J.; Sawyer, Richard L.
TITLE Estimating Measures of Pass-Fail Reliability from Parallel Half-Tests.
INSTITUTION American Coll. Testing Program, Iowa City, IA. Research Div.
PUB DATE 18 Jan 88
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); Licensing Examinations (Professions); *Pass Fail Grading; *Scores; Statistical Analysis; Test Construction; *Test Reliability
IDENTIFIERS Parallel Test Forms; *Spearman Brown Formula

ABSTRACT

Two methods for estimating measures of pass-fail reliability are derived, by which both theta and kappa may be estimated from a single test administration. The methods require only a single test administration and are computationally simple. Both are based on the Spearman-Brown formula for estimating stepped-up reliability. The non-distributional method requires only that the test be divisible into parallel half-tests; the normal method makes the additional assumption of normally distributed test scores. Bias for the two procedures is investigated by simulation, using a Monte Carlo study. For nearly normal test score distributions, the normal method performs slightly better than does the non-distributional method, but for moderately to severely skewed or symmetric platykurtic test score distributions, the non-distributional method is superior. Test results from a licensure examination are tabulated to illustrate the methods. (Author/SLD)

Revised Submitted Draft
1/18/88



# Estimating Measures of Pass-Fail Reliability from Parallel Half-Tests

David J. Woodruff, Research Psychometrician

and

Richard L. Sawyer, Director, Research and Statistical Services



The American College Testing Program
P.O. Box 168
Iowa City, IA 52243



Running Head: ESTIMATING PASS-FAIL RELIABILITY

# Estimating Measures of Pass-Fail Reliability from Parallel Half-Tests

Abstract

Two methods for estimating measures of pass-fail reliability are derived. The methods require only a single test administration and are computationally simple. Both are based on the Spearman-Brown formula for estimating stepped-up reliability. The non-distributional method requires only that the test be divisible into parallel half-tests; the normal method makes the additional assumption of normally distributed test scores. Bias for the two procedures is investigated by simulation. For nearly normal test score distributions, the normal method performs slightly better than the non-distributional method, but for moderately to severely skewed or symmetric platykurtic test score distributions the non-distributional method is superior. Test results from a licensure examination are used to illustrate the methods.

KEY WORDS: Cohen's kappa, licensure examination, pass-fail reliability, Spearman-Brown formula.

## Introduction

A primary component of the Standards for Educational and Psychological Testing (APA, 1985) with respect to licensure and certification examinations requires test publishers to report the reliability of pass-fail decisions (hereafter referred to as PF reliability). Hambleton and Novick (1974) proposed $\theta$, the proportion of consistently classified examinees, as a measure of PF reliability. Swaminathan, Hambleton, and Algina (1974) suggested that Cohen's (1960) kappa coefficient, denoted by $\kappa$, be used in place of $\theta$. Coefficient $\kappa$ is the proportion of consistently classified examinees, corrected for chance. Though it is commonly thought of as a measure of association rather than of agreement, $\phi$, the Pearson correlation between two dichotomous variables, equals $\kappa$ under certain circumstances that will be discussed. Thus, $\phi$ may also be used as a measure of PF reliability.

If two parallel test forms are available for administration to the same sample of examinees, then estimates for $\theta$ and $\kappa$ are easily obtained by the method of moments. If only one form of the test may be administered, then obtaining estimates for $\theta$ and $\kappa$ becomes much more difficult, both theoretically and computationally. Huynh (1976) developed a procedure for estimating $\theta$ and $\kappa$ which is based on a beta-binomial model and requires only one test administration. The computations involved are quite intricate, but Huynh (1976) also suggested a simpler method based on a normal approximation. Peng and Subkoviak (1980) further simplified Huynh's (1976) approximate method and presented evidence suggesting that their simplified procedure is superior to Huynh's. Brennan (1981) supplied tables which make the computations for Peng and Subkoviak's (1980) procedure relatively

simple. Subkoviak (1980) discussed several other methods for estimating $\theta$ and $\kappa$ when only one test form is available.

The purpose of this paper is to derive and illustrate two theoretically and computationally simple methods by which both $\theta$ and $\kappa$ may be estimated from a single test administration, when the test is divisable into parallel half-tests. One of the methods is based on normal theory; the other makes only minimal distributional assumptions. Bias for the two procedures is evaluated under a variety of test score distributions and test reliabilities using simulation techniques.

## Derivation of the Methods

Let X denote the total test and Y1 and Y2 the parallel half-tests (Lord and Novick, 1968) into which X is divisible. As will be seen later, the statistical assumptions defining parallelism for Y1 and Y2 may be relaxed so long as Y1 and Y2 are parallel (homogeneous) in content. Let A denote the dichotomous variable that equals 0 when an examinee fails X and equals 1 when an examinee passes X. The dichotomous variables B1 and B2 are similarly defined for Y1 and Y2. The three variables A, B1, and B2 require that passing scores be set for X, Y1, and Y2. The passing score for X is usually determined, at least in part, from criterion information distinct from the pass rate. It is assumed, however, that the passing scores for Y1 and Y2 are determined so that the pass rates for Y1 and Y2 are identical to that of X.

The proportion parameters describing the variables B1 and B2 may be expressed in the usual format of a 2 by 2 table as follows:

| Value of B1 | Value of B2: 0 | 1 | Total |
|---|---|---|---|
| 0 | $\pi_{00}$ | $\pi_{01}$ | q |
| 1 | $\pi_{10}$ | $\pi_{11}$ | p |
| Total | q | p | 1 |

In this table $\pi_{00}$ is the proportion of examinees in the population of interest who fail both Y1 and Y2, and $\pi_{01}$, $\pi_{10}$, and $\pi_{11}$ are defined analogously. Because the pass rate (p) is the same for B1 and B2, $\pi_{01} = \pi_{10}$. In terms of these proportion parameters,

$$\phi = 1 - \pi_{01}/(pq) , \tag{1}$$

$$\theta = \pi_{00} + \pi_{11} , \text{ and} \tag{2}$$

$$\kappa = (\theta - \theta_I)/(1 - \theta_I) , \tag{3}$$

where $\theta_I = p^2 + q^2$. The parameter $\theta_I$ is the value of $\theta$ when B1 and B2 are statistically independent. Given the above assumption that the pass rate is the same for B1 and B2, one can show that $\kappa = \phi$, as first noted by Cohen (1960) who also stated that $\kappa$ and $\phi$ are nearly identical so long as the pass rates differ by no more than .10.

Estimates for $\phi$, $\theta$, and $\kappa$ obtained by substituting observed proportions into (1), (2), and (3) would pertain to decisions based on the half-tests Y1 and Y2 and not, as is desired, on the whole test X. Let $Y1^*$ and $Y2^*$ be the doubled in length versions of Y1 and Y2, when the lengthening is in accordance

with the model of parallel measurements. Thus, $Y1^*$ and $Y2^*$ are parallel forms of X. Let $B1^*$ and $B2^*$ be the dichotomization of $Y1^*$ and $Y2^*$ under the asumption that the passing scores for $Y1^*$ and $Y2^*$ are chosen so that the pass rates are the same as for Y1, Y2, and X. Finally, let $\phi^* = \kappa^*$ and $\theta^*$ be the PF reliability coefficients corresponding to $B1^*$ and $B2^*$ (and, consequently, to A). Expressions for these coefficients will be derived below.

Calculations Using Normal Theory

One simple way to estimate $\phi^*$ and $\theta^*$ is a straightforward modification of the Huynh-Peng-Subkoviak (HPS) procedure to fit the present model of parallel half-tests. One can drop the HPS beta-binomial model assumption for item sampling, but keep the bivariate normal approximation for test scores. Let $K_q = (K - \mu_x)/\sigma_x$, where K is the passing score on the total test, and $\mu_x$ and $\sigma_x$ are the population mean and standard deviation for the total test. Under the assumptions of the model, $q = P[Z \leq K_q]$, where Z is a standard normal random variable, and $p = 1 - q$. Furthermore

$$\pi_{oo}^* = P[Z_1 \leq K_q, Z_2 \leq K_q; \rho^*] \qquad (4)$$

where $Z_1$ and $Z_2$ have a standard bivariate normal distribution with correlation coefficient $\rho^*$.

To estimate $K_q$, one can replace $\mu_x$ and $\sigma_x$ with their corresponding sample estimators. From the parallel half-scores Y1 and Y2, one can estimate the half-test reliability, $\rho$, then step it up to an estimate, $r_{SB}$, of the full-test reliability, $\rho^*$, using the Spearman-Brown formula. The estimates for $K_q$ and $\rho^*$ can then be substituted in Equation 4 to estimate $\pi_{oo}^*$. The tables in Brennan (1981) may be used to look up values for the estimate of $\pi_{oo}^*$, as may other tables of the bivariate normal distribution function.

Estimates for the probabilities $\pi^*_{01}$ and $\pi^*_{11}$ can then be computed from the relationships $\pi^*_{00} + \pi^*_{01} = q$ and $\pi^*_{11} + \pi^*_{01} = p$. Since $\theta^* = \pi^*_{00} + \pi^*_{11}$ and $\phi^* = \kappa^* = 1 - \pi^*_{01}/(pq)$, the normal model estimates for $\pi^*_{00}$, $\pi^*_{11}$, and $\pi^*_{01}$ can be immediately converted to estimates for $\theta^*$ and $\phi^*$.

The practical difference between the method proposed here and the HPS method is the use of the stepped-up reliability estimate, $r_{SB}$, in place of KR21. The basic theoretical difference is that the method proposed here relies on a parallel half-test model rather than on a beta-binomial model for the full test.

A Non-Distributional Method

In the normal theory model above, the half-length reliability, $\rho$, was stepped up to the full-length reliability, $\rho^*$, by the Spearman-Brown formula. Alternatively, one could step up the half-length PF reliability, $\phi$, directly: $\phi^*_{SB} = 2\phi/(1+\phi)$. Substituting the expression for $\phi$ given in (1) into this expression for $\phi^*_{SB}$ and simplifying yields

$$\phi^*_{SB} = 1 - \frac{\pi_{01}}{2pq - \pi_{01}} \tag{5}$$

where $p = \pi_{01} + \pi_{11}$ is the pass rate and $q = 1-p$. Because $\phi^* = 1 - \pi^*_{01}/(pq)$, it follows that

$$\phi^*_{SB} - \phi^* = \frac{\pi^*_{01}}{pq} - \frac{\pi_{01}}{2pq - \pi_{01}} . \tag{6}$$

(Note that the pass rate $p = \pi_{01} + \pi_{11} = \pi^*_{01} + \pi^*_{11}$ is the same for both the half-length and the full-length tests.) For non-zero $\pi_{01}$, the left side difference in (6) is zero if and only if $\pi^*_{01} = [1/(1 + \phi)]\pi_{01}$. However,

because $0 \leq \pi^*_{01} \leq \pi_{01}$, the first term of the right side difference is greater than 0 and less than $\pi_{01}/(pq)$, and this leads to the following upper and lower bounds for the left side difference in (6):

$$-\frac{\pi_{01}}{2pq - \pi_{01}} \leq \phi^*_{SB} - \phi^* \leq \frac{\pi_{01}}{pq} - \frac{\pi_{01}}{2pq - \pi_{01}} . \qquad (7)$$

Since each side of this inequality approaches 0 as $\pi_{01}$ approaches 0, $\phi^*_{SB}$ becomes a better approximation to $\phi^*$ as the half-test reliability increases, though, as will be argued, $\phi^*_{SB}$ can be a useful approximation to $\phi^*$ when the half-test reliability is only moderate.

Technically, $\phi^*_{SB}$ is the reliability of a test composed of the two dichotomously scored parts B1 and B2 or, equivalently, the correlation between two parallel forms of such a test. These test scores would be trichotmous variables taking the values 0, 1, and 2. Consequently, the interpretation of $\phi^*_{SB}$ as the correlation between $B1^*$ and $B2^*$ is an approximation, because $B1^*$ and $B2^*$ are dichotomous variables.

More specifically, let C = B1 + B2 and C′ = B1′ + B2′ where the prime denotes a parallel measurement. The coefficient $\phi^*_{SB}$ equals the correlation between the two parallel measurements, C and C′, both of which may take the values of 0, 1, or 2. For the measurement C, the value 0 occurs if the examinee fails both B1 and B2. The value 1 occurs if the examinee pass one but fails the other. If the examinee passes both B1 and B2, then C takes the value of 2. The values of C′ are similarly defined in terms of B1′ and B2′.

If B1 and B2 are reasonably reliable, then there should be relatively few examinees with scores of 1 on C and C′. Assume that the group of examinees scoring 1 on C is approximately the same as the group of examinees with scores of 1 on C′. Let the dichotomous variables D and D′ be defined by dividing

this group of examinees in half and assigning one half scores of 0 and the other half scores of 2.

The dichotomous variables D and D′ will have approximately the same means as C and C′, but their variances will be larger. The covariance between D and D′ will also be larger than the covariance between C and C′. Consequently, the correlation between D and D′ should be approximately equal to the correlation between C and C′, which is $\phi^*_{SB}$. However, since D and D′ may be viewed as a dichotomous grouping of the trichotomous variables C and C′, $\phi^*_{SB}$ may be close to but slightly greater than the correlation between D and D′ because of attenuation due to grouping. The variables D and D′ take the values of 0 or 2 while the variables $B1^*$ and $B2^*$ take the values of 0 or 1. While the variables D and D′ are defined differently than the variables $B1^*$ and $B2^*$, their values are linearly related with a slope equal to two and an intercept equal to zero. Insofar as the variables D and D′ serve as approximate representations for the variables $B1^*$ and $B2^*$, on a different scale, $\phi^*_{SB}$ may be interpreted as an approximation to $\phi^*$, the correlation between $B1^*$ and $B2^*$.

The approximation $\phi^*_{SB}$ is related to $\theta^*$ as follows. First, algebraic manipulations show that $\pi^*_{11} = pq\phi^* + p^2$ and $\pi^*_{00} = pq\phi^* + q^2$. Combining these equations with the relationship $\theta^* = \pi^*_{00} + \pi^*_{11}$ results in

$$\theta^*_{SB} = 2pq\phi^*_{SB} + p^2 + q^2 . \qquad (8)$$

The relationship to $\kappa^*$ is more simple: $\kappa^*_{SB} = \phi^*_{SB}$.

In practice, it may not be possible to construct exactly parallel half-tests from a single test, and as a result, $\pi_{01}$ may not be equal to $\pi_{10}$. If the difference in the observed proportions $p_{01}$ and $p_{10}$ corresponding

to $\pi_{01}$ and $\pi_{10}$ are minor in relation to the sample size, then $p_{01}$ and $p_{10}$ may be replaced by their average, and the marginal proportions adjusted accordingly. The above formulas are then applied to the modified 2 x 2 table of observed proportions.

## Simulation Results

A Monte Carlo investigation was undertaken to evaluate the accuracy of the non-distributional procedure as well as the normal procedure. An important application of PF reliability indices is in the area of certification and licensure examination where there is usually at least several hundred and often times many thousand examinees taking the test. Hence, it was considered to be more important to investigate the bias of the procedures for large sample sizes rather than to compare the small sample standard errors of the procedures.

The present simulations were undertaken on an IBM 4381 mainframe using SAS version 5 (SAS Institute Inc., 1985), except that IMSL (IMSL, 1987) function BNRDF was used to evaluate the bivariate normal cumulative distribution function for the normal method. Six simulation situations were considered. Within each of three different test score distribution shapes: nearly normal, platykurtic, and negatively skewed, two full-test reliabilities were considered: .92 and .71, where the full-test reliability is defined as the Spearman-Brown stepped-up correlation between the half-tests. For each of the six simulation situations, two replications were done where one replication consists of the generation of four half-test scores for 20,000 examinees under the following model:

$$Y_{ij} = \gamma^2 T_i + \gamma E_{ij} + \beta \qquad i = 1, \ldots, 20{,}000 \text{ and } j = 1, 2, 3, 4,$$

where $\gamma$ and $\beta$ are parameters and T and all $\{E_{ij}\}$ are independent variates generated from the standard normal distribution. All half-test scores were

rounded to integer values, and the full-test scores were computed as $X_1 = Y_1 + Y_2$ and $X_2 = Y_3 + Y_4$. Various degrees of symmetrical and asymmetrical truncation on T and to a much lesser extent symmetrical .cation on the E's was used to control the distributional shapes of the test scores being generated. Formulas for the mean and variance of truncated normal variables are available (Johnson & Kotz, 1970), and these in combination with the $\gamma$ and $\beta$ parameters permitted some control over the means, variances, and reliabilities of the test scores.

Full-test characteristics for the six simulation situations are presented in Table 1. These situations were chosen as representative of those

---

Insert Table 1 about here

---

encountered in practice. Test score distributions for licensure, certification, and various other selection examinations are frequently but not exclusively found to be negatively skewed as is illustrated by an example presented later. Alternatively, Lord (1955) found for professionally constructed educational examinations that if they were symmetric they tended to be platykurtic (flat). Lord's (1955) finding was reaffirmed with a random sample of 40,000 examinees from a recent administration of the ACT Assessment examination. The distributions of raw scores for the four subtests comprising the ACT Assessment were approximately symmetric with skews ranging from -.2 to +.25 but with kurtoses ranging from -.40 to -.96. Finally, since test scores are inherently bounded, rather than consider an exactly normal distribution situation, a nearly normal situation with slight platykurtosis was used instead. Though full-test reliabilities for professionally constructed examinations are usually in the nineties or at least the eighties, subtest

reliabilities may be lower and so both high and low reliabilities represented by .92 and .71 were considered.

Failure rates of ten and thirty percent were selected for investigation as they seemed to represent a realistic range. Due to the integer nature of the generated test scores, it was not always possible to achieve exactly ten or thirty percent failure rates for the full-tests. Rather, the failure rates ranged from 8.5% to 11% and from 29% to 32.5% across the six situations.

For an estimator T of some parameter $\psi$, bias is defined as $E(T) - \psi$. The parameters of interest are the PF reliability indices $\theta$ and $\phi = \kappa$. For the distributions modeled, the theoretical values of these parameters are not known. However, the simulations included the generation of two full-test scores for all simulated examinees, and consistent estimates for the parameters were obtained by applying the method of moments (MOM) to the 2 by 2 table derived from the pairs of full-test scores. With a N of 20,000 these consistent MOM estimates should, for practical purposes, accurately reflect the true parameter values. In what follows, these estimates are denoted by a carat, but have no subscript. The two estimation methods compared were the normal and non-distributional methods both of which are computed for the first full-test score $X_1 = Y_1 + Y_2$ only. Normal method estimates have N as a subscript while the non-distributional method estimates have SB as a subscript. The bias for each method is estimated as the difference between its estimate and the consistent MOM estimate. This approach of using the MOM estimate as the parameter is similar to that used by Peng and Subkoviak (1980), Huynh and Sanders (1980), and Subkoviak (1978).

Table 2 presents MOM estimates for the PF reliability indices $\theta$ and $\phi = \kappa$ as well as estimated biases for the normal and non-distributional methods. Results are presented for two replications under all

six simulation situations and for approximate fail rates of ten and thirty percent.

---

Insert Table 2 about here

---

The replications in Table 2 reveal some variability in the bias estimates even with an N of 20,000. Despite this variability, clear patterns do emerge. Focusing first on $\theta$, it can be seen that for the two nearly normal situations the normal method is never significantly worse and in one case appreciably better than the non-distributional method, though the bias for both methods is modest. With the four non-normal situations, the pattern is reversed. The non-distributional method is never substantially worse and usually considerably better than the normal method, but again, both methods usually show only modest bias.

Turning next to $\phi$, the pattern is similar but the biases are generally larger, the latter result having also been observed by Peng and Subkoviak (1980) and Huynh and Sanders (1980) with their methods. For the two nearly normal situations, the normal method is appreciably better than the non-distributional method though the latter method performs reasonably well. With the four non-normal situations, the non-distributional method usually performs fairly well and is considerably better than the normal method which has rather large bias when the fail rate is 10%.

It is interesting to note that while the normal method sometimes yields positive and sometimes negative bias estimates, the biases in Table 2 are always positive for the non-distributional method. In the derivation of the non-distributional method, it was suggested that insofar as the method may be biased, the bias would be positve and attributable to attenuation due to

grouping. It is also worthwhile to note that previous simulation studies by Peng and Subkoviak (1980) and Huynh and Sanders (1980) found that the HPS method and Huynh's beta-binomial method had biases similar in magnitude to those found for the present methods, though the previous studies concentrated on short tests while the focus of the present study is long tests. However, Huynh's beta-binomial model was applied to the current simulated data, but the results are not reported because, as was expected, its performance was very similar to the performance of the normal method. (In applying the beta-binomial method, the number of items on each test was chosen so that KR21 was close in value to $\rho_{SB}$ with the constraint that test length could never be less than the maximum observed score.)

In summary, neither method shows large bias when estimating $\theta$, though the normal method generally shows less bias than the non-distributional method when the test scores are approximately normally distributed while the opposite holds when they are not. In estimating $\phi$, the non-distributional method generally shows mild to moderate positve bias, but is considerably less biased than the normal method when the test scores are not normally distributed with the reverse being true when they are. These results indicate that when the sample size is large and the test score distribution shows substantial departures from normality the non-distributional method should yield more accurate estimates of $\theta$ and especially $\phi$ than the normal method.

In the next section, the methods are illustrated with data from a licensure examination.

## An Illustrative Example

The data used here are from a licensure examination containing 300 scored items. The test is divided into two separately timed parts consisting of 150 scored items each. The two parts were constructed to be equally difficult

based on field test data and were matched in content according to the test's table of specification. A group of approximately twenty expert judges rated the 300 scored items using the Angoff (1971) method. The judges also rated what proportion of items a minimally competent examinee should answer correctly in each of the many content areas covered by the test. A passing score for the total test of at least 200 items correct was determined from a weighted average of the judges' item and area ratings.

The method requires that passing scores be determined for the two parts. From a strictly statistical perspective, these passing scores should be chosen so that the passing rates on the two parts are equal to each other and to the percentage passing on the full test. If a representative sample of examinees is available, then the half-test passing scores may be determined solely from the passing rates. The half-test passing scores need not be taken as one-half of the full test passing score, nor do the half-test passing scores have to sum to the full test passing score. In general, these last two conditions will not be fulfilled when the half-test passing scores are determined by equating the passing rates.

In many applications, it may be possible to integrate psychometric and statistical considerations. If criterion data, such as expert judges' ratings, are available, then these data may also be employed in determining the half-test passing scores. Consider the present example: In rating the items and areas, the judges' only concern was to establish a passing score for the total test which would determine whether or not an examinee should be licensed. However, using the same weights as were used to determine the passing score for the total test, the item and area ratings were also used to determine passing scores for the two parts. Both parts received 100 items correct as passing scores based on the expert judges' ratings. After the test

was administered and the results analyzed, these passing scores were changed to 102 for part one and 98 for part two. The reason for the change was that in the total group of examinees the average score for part one was approximately four points higher than that for part two, and with these adjusted passing scores the part one and part two passing rates were nearly identical to each other and to the full test passing rate. In this example, empirical results from a large representative sample were used to adjust the judges' ratings. Note that no decisions about examinees were based on the part one and part two passing scores. Their only function is in estimating the full test PF reliability. The fact that the half-test passing scores sum to the full test passing score is due to the long length and corresponding high reliability of the two parts. If the two parts were shorter, this condition would likely be violated.

Summary statistics for the total group and selected subgroups of examinees are presented in Table 3.

---

Insert Table 3 about here

---

The subgroup data are presented to illustrate the method for different sample sizes and for groups with different observed passing rates. The determination of the subgroups is based upon whether an examinee was taking the test for the first time or was repeating the test; and whether an examinee graduated from an accredited or nonaccredited university.

The sample alpha coefficients are derived from scores on the examination's five subtests which differ in content, rather than from the item scores. This is why the sample alphas are slightly smaller than the sample KR21's. Since the subtests differed widely in length, average subtest scores

(rather than total subtest scores) were used for computing the alphas.

One would generally expect that the stepped-up reliability coefficient, $r_{SB} = 2r(Y1,Y2)/[1 + r(Y1,Y2)]$, would be larger than KR21, though that is not always the case in Table 3. Most likely, this is due to the long length of the test. In any case, the two reliability coefficients are very similar in all subgroups.

The data in Table 3 show that while Y1 and Y2 have similar standard deviations, their means tend to differ; hence Y1 and Y2 are not precisely parallel. Moreover, the negative skewness coefficients suggest moderate to severe departures of the data from normal distributions.

The stepped-up phi coefficient, $\phi^*_{SB}$, is based, though, on neither of these assumptions, but on the assumption that B1, B2, and A have the same pass rates. An indication of how well the data satisfy this assumption can be found in the pass rate column of Table 3. The passing scores for Y1 and Y2 were chosen so that the assumption would be fulfilled in the total group of examinees. The assumption continues to be met in the group of accredited first-time examinees but is violated to varying degrees in the remaining three groups. As was previously discussed, the observed proportions may be smoothed in the application of the non-distributional method. The "smoothed half-test proportions" in Table 4 were obtained by replacing the two off-diagonal proportions with their average. The estimated full-test proportions and PF reliability indices in Tabel 4 were computed from the smoothed proportions. Though the sample proportions in Table 4 are reported to only 3 digits, the computations for the PF indices used 4 digits.

---

Insert Table 4 about here

---

The HPS estimates of the full-length reliability indices are based on Brennan's (1981) tables. Because, as was noted above, KR21 is nearly identical to the stepped-up reliability coefficient, $r_{SB}$, the HPS reliability indices are nearly identical to those that result from applying the normal model to the half-test data, as discussed earlier in this paper. For this reason, the PF reliability indices associated with the normal model are omitted from Table 4.

Comparing the SB and HPS estimates in Table 4 shows that they yield nearly identical estimates for $\theta^*$, but that their estimates for $\kappa^*$ are sometimes discrepant. The results from this example are consistent with the simulation results. When N is large and the test score distribution is substantially skewed, as in the first two groups in Table 4, the two methods give substantial different estimates for $\phi = \kappa$. The simulation results indicate that the SB method estimates should be more accurate, and this is supported in this example by the similarity of the HPS estimates to the unstepped-up half-test MOM estimates of $\phi$. Doubling the length of the test should increase $\phi$ at least a moderate amount; that was always the case in the simulations. The other three groups are considerably less skewed (with fairly normal kurtoses also) and here the HPS and SB estimates for $\phi$ are more similar and substantially increased over the unstepped-up half-test MOM estimates of $\phi$.

## Summary and Discussion

The methods for computing PF reliability presented in this paper require only one test administration and use the Spearman-Brown formula to obtain stepped-up estimates of PF reliability which are computed from parallel half-tests. They thus require that the test be divisible into two parts that are equivalent in their content and approximately equivalent in certain

statistical characteristics. If this is not the case, then one of the beta-binomial model based methods discussed by Subkoviak (1980) could be used such as the one by Huynh (1976). However, the beta-binomial method is computationally complex and appears more appropriate when tests are short and homogeneous in content and item difficulties. For long tests which are heterogeneous in content and item difficulties, such as licensure examinations, the Peng and Subkoviak (1980) approximation should yield results nearly identical to those from the beta-binomial method. Brennan (1981) presents tables which make the Peng and Subkoviak computations relatively simple. Brennan (1981) also discusses other PF reliability indices in addition to $\theta$ and $\kappa$ .

If the test is divisible into parallel half-tests, then the methods derived within possess certain advantages. Instead of KR21, which is used in the HPS method, the normal method uses a Spearman-Brown stepped-up half-tests intercorrelation as an estimate of the correlation between two full-tests. This latter estimate is based on less restrictive assumptions than KR21, and as a consequence the normal method has wider applicability than the HPS method. In particular, it should be better suited to long heterogeneous tests such as licensure examinations, though this may not always be the case as is illustrated by the example given within. Of more importance, however, is the non-distributional method which discards distributional assumptions altogether. The simulation results support the conclusion that when N is large and the test score distribution is non-normal, the non-distributional method will yield more accurate estimates than the normal method especially for $\phi = \kappa$ and especially for smaller (.10) failure rates.

Though the non-distributional method outperformed the normal method when normality was violated, it still displayed mild to moderate bias. The bias,

however, was always positive, in contrast to the normal method, and this suggests that it may be worthwhile to investigate strategies for correcting the bias. Also, the magnitude of the biases found here were generally similar to those found by Peng and Subkoviak (1980) for their approximate method and to those found by Huynh and Sanders (1980) for the beta-binomial method, though these two studies focused on short tests. Finally, the simulation results obtained here are only applicable when sample sizes are large. An investigation of the behavior of the non-distributional method when sample size is small, test length is short, and test score distributions are non-normal could extend the applicability of the method to situations other than the one of professional licensure and certification examinations which was considered here.

## References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 508-600.

Brennan, R. L. (1981). Some statistical procedures for domain-referenced testing: A handbook for practitioners. ACT Technical Bulletin No. 38. Iowa City, IA: The American College Testing Program.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Committee of AERA, APA, & NCME to Develop Standards. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Hambleton, R. K. & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Huynh, H. (1976). On the reliability of decisions in domain referenced testing. Journal of Educational Measurement, 13, 253-264.

Huynh, H. & Sanders, J. C. (1980) Accuracy of two procedures for estimating reliability of mastery tests. Journal of Educational Measurement, 17, 351-358.

IMSL. (1987). STAT/LIBRARY: FORTRAN subroutines for statistical analysis (Vol 3). Houston: Author.

Johnson, N. L. & Kotz, S. (1970). Distributions in statistics: Continuous univariate distributions -1. Boston: Houghton Mifflin.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. Educational and Psychological Measurement, 15, 383-389.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Peng, C. J. & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 17, 359-368.

SAS Institute Inc. (1985). SAS user's guide basics, version 5 edition. Cary, NC: Author.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15, 111-116.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A Guide to Criterion-Referenced Test Construction. The John Hopkins University Press: Baltimore, MD, 267-291.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 11, 263-267.

Table 1. Characteristics of the Simulation Distributions

| Dist. | Rel. | Mean | S.D. | Skew | Kurtosis |
|---|---|---|---|---|---|
| Normal* | .92 | 160 | 23.0 | .01 | -.17 |
| Normal* | .71 | 70 | 8.5 | .00 | -.14 |
| Flat | .92 | 140 | 25.4 | .01 | -.93 |
| Flat | .71 | 50 | 8.7 | -.01 | -.65 |
| Skewed | .92 | 165 | 20.5 | -.84 | .47 |
| Skewed | .71 | 57 | 5.5 | -.63 | .42 |

*These are nearly normal with slight platykurtosis.

Table 2: Parameter Estimates and Bias Estimates for the Pass-Fail Reliability Indices -- N=20,000 and $\phi^* = \kappa^*$

| Dist. | $\rho_{SB}$ | Sim. | 10% Failing | | | 30% Failing | | | 10% Failing | | | 30% Failing | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\theta}^*$ | $\hat{\theta}^*_{SB}-\hat{\theta}^*$ | $\hat{\theta}^*_{N}-\hat{\theta}^*$ | $\hat{\theta}^*$ | $\hat{\theta}^*_{SB}-\hat{\theta}^*$ | $\hat{\theta}^*_{N}-\hat{\theta}^*$ | $\hat{\phi}^*$ | $\hat{\phi}^*_{SB}-\hat{\phi}^*$ | $\hat{\phi}^*_{N}-\hat{\phi}^*$ | $\hat{\phi}^*$ | $\hat{\phi}^*_{SB}-\hat{\phi}^*$ | $\hat{\phi}^*_{N}-\hat{\phi}^*$ |
| Nearly Normal | .92 | 1 | .94 | .007 | -.003 | .89 | .017 | .000 | .70 | .034 | -.009 | .73 | .040 | .001 |
| | | 2 | .95 | .007 | -.004 | .88 | .021 | .003 | .71 | .029 | -.018 | .72 | .051 | .008 |
| Nearly Normal | .71 | 1 | .90 | .001 | -.006 | .78 | .008 | -.005 | .41 | .004 | .006 | 48 | .026 | .007 |
| | | 2 | .90 | .004 | -.001 | .79 | .013 | -.009 | .40 | .011 | .016 | .50 | .035 | -.006 |
| Flat | .92 | 1 | .92 | .012 | .028 | .90 | .015 | -.012 | .55 | .086 | .133 | .77 | .037 | -.043 |
| | | 2 | .92 | .009 | .026 | .91 | .016 | -.014 | .57 | .066 | .114 | .78 | .044 | -.050 |
| Flat | .71 | 1 | .88 | .008 | .018 | .79 | .008 | -.015 | .31 | .053 | .108 | .53 | .012 | -.037 |
| | | 2 | .87 | .012 | .021 | .79 | .003 | -.014 | .29 | .088 | .124 | .52 | .000 | -.035 |
| Skewed | .92 | 1 | .96 | .009 | -.007 | .91 | .016 | -.030 | .77 | .041 | -.089 | .78 | .038 | -.047 |
| | | 2 | .96 | .010 | -.008 | .91 | .013 | -.029 | .77 | .049 | -.088 | .78 | .031 | -.047 |
| Skewed | .71 | 1 | .93 | .005 | -.022 | .80 | .013 | -.044 | .56 | .057 | -.147 | .55 | .031 | -.051 |
| | | 2 | .93 | .000 | -.026 | .80 | .012 | -.037 | .57 | .030 | -.156 | .54 | .027 | -.038 |

Table 3. Summary Statistics and Reliability Coefficients, by Examinee Group

| Examinee group | N | Summary statistics | | | | | Reliability coefficients | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Variable | Mean | SD | Skewness | Pass rate | Alpha(X) | KR21(X) | r(Y1,Y2) | $r_{SB}$ |
| Total group | 4828 | X | 236.06 | 29.78 | -1.41 | .900 | .92 | .95 | .90 | .94 |
| | | Y1 | 119.93 | 15.11 | -1.48 | .894 | | | | |
| | | Y2 | 116.13 | 15.48 | -1.24 | .897 | | | | |
| Accredited first-time | 3999 | X | 241.96 | 22.59 | -0.80 | .956 | .87 | .91 | .84 | .91 |
| | | Y1 | 122.97 | 11.39 | -0.88 | .952 | | | | |
| | | Y2 | 118.99 | 12.20 | -0.69 | .949 | | | | |
| Accredited repeating | 548 | X | 224.01 | 28.73 | -0.27 | .812 | .90 | .93 | .88 | .94 |
| | | Y1 | 113.59 | 14.79 | -0.39 | .790 | | | | |
| | | Y2 | 110.42 | 14.84 | -0.17 | .821 | | | | |
| Non-accredited first-time | 94 | X | 187.22 | 49.61 | -0.14 | .404 | .96 | .98 | .94 | .97 |
| | | Y1 | 95.01 | 24.28 | -0.19 | .436 | | | | |
| | | Y2 | 92.21 | 26.06 | -0.17 | .436 | | | | |
| Non-accredited repeating | 187 | X | 169.74 | 39.76 | -0.03 | .209 | .91 | .96 | .90 | .95 |
| | | Y1 | 86.05 | 20.38 | -0.13 | .198 | | | | |
| | | Y2 | 83.69 | 20.38 | 0.01 | .251 | | | | |

Table 4: Pass-Fail Proportions and Reliability Indices, by Examinee Group

| Examinee group | N | PF decision B1 | PF decision B2 | Half-test proportion Raw | Half-test proportion Smoothed | Full-test Proportion | PF reliability indices: Half test $\hat{\phi}=\hat{\kappa}$ | PF reliability indices: Half test $\hat{\theta}$ | PF reliability indices: Full-test $\hat{\phi}^*_{SB}=\hat{\kappa}^*_{SB}$ | PF reliability indices: Full-test $\hat{\theta}^*_{SB}$ | PF reliability indices: Full-test $\hat{\phi}^*_{HPS}=\hat{\kappa}^*_{HPS}$ | PF reliability indices: Full-test $\hat{\theta}^*_{HPS}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total group | 4828 | 0 | 0 | .078 | .078 | .089 | .72 | .95 | .84 | .97 | .76 | .95 |
| | | 0 | 1 | .028 | .026 | .015 | | | | | | |
| | | 1 | 0 | .025 | .026 | .015 | | | | | | |
| | | 1 | 1 | .870 | .870 | .881 | | | | | | |
| Accredited first-time | 3999 | 0 | 0 | .030 | .030 | .037 | .59 | .96 | .74 | .98 | .61 | .98 |
| | | 0 | 1 | .018 | .019 | .012 | | | | | | |
| | | 1 | 0 | .021 | .019 | .012 | | | | | | |
| | | 1 | 1 | .931 | .931 | .938 | | | | | | |
| Accredited repeating | 548 | 0 | 0 | .133 | .133 | .156 | .61 | .88 | .76 | .92 | .76 | .92 |
| | | 0 | 1 | .077 | .061 | .038 | | | | | | |
| | | 1 | 0 | .046 | .061 | .038 | | | | | | |
| | | 1 | 1 | .745 | .745 | .768 | | | | | | |
| Non-accredited first-time | 94 | 0 | 0 | .500 | .500 | .527 | .74 | .87 | .85 | .93 | .90 | .95 |
| | | 0 | 1 | .064 | .064 | .037 | | | | | | |
| | | 1 | 0 | .064 | .064 | .037 | | | | | | |
| | | 1 | 1 | .372 | .372 | .399 | | | | | | |
| Non-accredited repeating | 187 | 0 | 0 | .722 | .722 | .744 | .69 | .89 | .82 | .94 | .78 | .92 |
| | | 0 | 1 | .080 | .054 | .032 | | | | | | |
| | | 1 | 0 | .027 | .054 | .032 | | | | | | |
| | | 1 | 1 | .171 | .171 | .193 | | | | | | |